What Archives want

— the requirements for digital technology

Richard Wright
Technology Manager, BBC Archives

As the world goes digital, archives are moving from their kilometres of shelves to a brave new world where the holdings are invisible, on some sort of IT system. Can this really happen? In the BBC we have 100 km of shelves in our main archive and about 3.5 million physical items (video and audio tapes, reels of film). The holdings are the permanent result of all that the BBC has meant and accomplished – because broadcasting, itself, simply goes out into the ether and disappears.

In the BBC archive we have decades of experience in dealing with this media. But how can we move all this content – this almost sacred legacy – into IT systems and still sleep at night? This article explains how it might be done.

This article reduces “what archives want” from IT storage to two simple issues:

- ***persistence*** – we want to get back what we put in;
- ***currency*** – we want to be able to use what we get back.

What we now want is for the storage industry to tell us the costs, and the cost trade-offs, for achieving *persistence* and *currency*.

Archivists are used to dealing with media and so, as we confront digital technology, we start off trying to gain a detailed understanding of storage technology and devices. But deeper study ultimately leads *away* from all such details about storage technology and into the realization that what matters is the *service* provided by the storage.

In the digital world, archives and storage are parting company. **Archivists** will manage the content, concentrating on the metadata (catalogue and other finding aids, rights data) required to manage the content – and defining the requirements for *storage services* (but NOT for storage systems and media). **Storage service providers** – using a range of technologies – will fulfil these requirements.

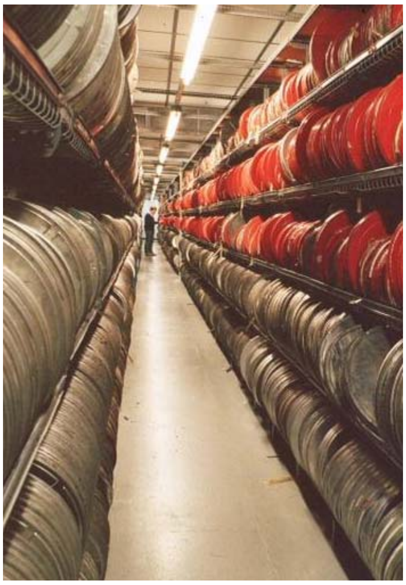

Figure 1
Part of the 100 km of shelves at BBC WIndmill Road

What archives want

What archives want from storage is a question that cannot be answered in isolation because the fundamental question is about what the archive needs and wants to do. So there is a range of requirements, at multiple levels:

Figure 2
Mass storage – 1960s style

Credit: Appaloosa

- The requirements of archives start with the **archive service requirements**: what an archive is good for; what an archive does.
- Below that are the **functional storage requirements**: the *function* the storage fulfils. Successful digital archives will use storage that functions adequately, storage that serves its purpose.
- Below that are **storage service requirements** – the technical requirements about gigabytes and bandwidth.
- Finally there are **storage media requirements**: how storage operates. The archive has to have some physical reality, somewhere. The contention of this article is that archivists (well, some of them) will initially be interested in digital media, but will quite quickly move back up in level to being primarily interested in the service, not the media.

Archives perform services – or they're of no use and they risk disappearing. As archives move from a *storehouse* to a *service provider* perspective, they move away from storage as a primary activity. Ultimately, archives and storage devices will part company. Digital archives will use a *storage service provider*, just as so many other IT functions now use service providers of one sort or another (ranging from networks to data centres). But digital archives need a service provider who understands archives and understands storage – particularly long-term storage – and it is surprisingly hard to get such expertise from the standard IT industry.

The two archives

The IT industry has trouble understanding the requirements of archives, because the IT industry already uses the word "archive" for something else)and so it doesn't expect to have to unlearn the definition of a word it already knows and uses – in order to learn what broadcasters mean by the word "archive".

Here are the main differences:

IT definition of *Archive*	Broadcasters definition of *Archive*
Where data goes to die: where data from an application goes when it is no longer needed by the application.	**Where data lives:** a repository, ready and waiting to be accessed.
Where data has to be restored (to the originating application) before it can be used: when data from an application has been "archived", any subsequent access requires a *restore* function – which typically takes system-manager intervention and a day's time, or even several days.	**Where data lives:** the archive *is* the application. The user of the archive expects to get to data within seconds, just like for any other application. However, the user of an audiovisual archive may be willing to wait longer for delivery of full-quality video. Providing catalogue data and browse-quality data is available within seconds, acceptable delivery times for full-quality video could be many minutes, or even an hour or so.
An archive is the place where data goes when it is no longer wanted.	**An archive is** the place where people go to get archive content.
An archive sits somewhere **behind an application.**	An archive is the application.

Archive requirements

What do archives want? Archive requirements in four areas are presented, beginning at the top with what archives do, and working down.

- **Archive service requirements** – what archives do for others;
- **Archive functional storage requirements** – what archives want from storage;
- **Storage service requirements** – what storage does for archives;
- **Storage media requirements** – how storage operates.

What archives do

Media archives in broadcasting exist mainly for the purpose of re-use of their content. Not everyone is aware of the high level of use of archive material in broadcasting: this is an article on *storage*, not on how wonderful archives are. But a few facts may set the perspective.

In the BBC, the archive provides about 30% of TV news – and this figure is rising with the introduction of server-based archive content. The BBC archive at Windmill Road in London responds to about 600 requests for content per day, and issues about 2000 items per day. Overall, about 20% of archive content is requested per year. As the BBC archive moves to direct public access, these figures could increase enormously. Direct public access to the physical archive – the shelves in Brentford – was impossible. **The major difference between a physical archive and a digital archive is not the storage – it's the access,** the potential for opening the archives to full access.

The specific functions for a digital archive are to provide the same (preferably, better) services, but using files on mass storage, distributed electronically, instead of on physical items taken from shelves and distributed by van and by hand.

All the above *archive service requirements* involve media held in computer files. There is a major issue with all file formats: **obsolescence**. Broadcast archives have been dealing with format obsolescence ever since ¼″ audio tape replaced gramophone recording, and videotape replaced film, and 1″ tape replaced 2″. Unfortunately in the digital archive the problem may well get worse rather than better.

What a digital archive needs to perform

A digital archive will need to perform the following operations (at least):

- **Acquisition**:
  - For new material: bring files into the digital archive;
  - Legacy material: digitization from physical items to files.
- **Documentation**:
  - An archive travels on its catalogue. As archives "go digital", the catalogue becomes the major value-added service of the archive.
- **Viewing**:
  - The archive will have to support a multiplicity of "proxies", because bandwidth will be insufficient to move high-resolution video files as quickly as would be the case with MPEG-4 (or whatever) viewing files.
  - Catalogue search, viewing and rough edit will, ideally, be combined in a single asset-management application.
- **Re-Use**:
  - Full-quality material will have to be delivered, as files, to edit suites or wherever else they are needed.
- **Asset management and life-cycle management**:
  - There is a set of birth-to-death processes here, based on processes established in the document management world (where they started "going digital" 20 years ago). Principal issues include access control, version control and digital rights management.

However digital broadcast archives will share this problem with all digital libraries, because file-format obsolescence affects absolutely all forms of digital files.

What broadcast engineers do not know, and what all-too-few IT experts know, is that the digital library world has weapons in the fight against digital obsolescence: **digital library process technology**.

Libraries have been developing their own technologies for decades. More than 25 years ago, libraries started putting their catalogues online (OPACs [1]), and coming up with standards to allow federated searching of multiple OPACs, using what became the NISO/ANSI and then ISO standard Z39.50 [2].

Starting roughly ten years ago, digital library technology has been looking at the issue of how to bring content into digital storage and get it out again reliably, and with sufficient knowledge about the content to be able to use it (read, view or listen to it; open it in a relevant application) – in perpetuity!

This work has led to the OAIS standard – the Open Archive Information System [3]. The OAIS standard provides detailed rules – processes – for moving material into a “trusted digital repository”, keeping it live while it’s there, and moving it out for use in a way that doesn’t compromise the integrity of the repository – and maximizes the chance that the material coming out will in fact be usable.

The OAIS work is extensive and complex, but that’s not the main problem. The main problem for those of us in audiovisual archives is that broadcasting doesn’t know anything about OAIS (or anything else in the digital library world, generally). This situation is a particular feature of the broadcast engineers (and IT staff) that “look after” the technical issues of archives within broadcasting.

My plea to technical people reading this article is: ask your archivists about archive technology! They may be able to tell you something.

What archives want from storage

Regarding how storage works, archives really only want information in two areas, as already stressed above:

- **Persistence** – the ability to get content out of storage;
- **Currency** – the ability to use that content.

These two terms are not standard in the storage industry, but they are basic concepts (under various labels) of digital preservation technology – and from the work on storage done under the SAM (Storage and Archive Management) part of EC project PrestoSpace [4]. I use these concepts, rather than standard storage terminology, because of the mismatch between the information that broadcast archives require, and the statistics generally available.

Persistence

Persistence is not a standard term in either archives or in the storage industry, although it is a standard IT term in the context of the Worldwide Web:

[1] Wikipedia: **http://en.wikipedia.org/wiki/OPAC**

[2] ISO standard Z39.50: **http://www.loc.gov/z3950/agency/**

[3] OAIS: **http://ssdoo.gsfc.nasa.gov/nost/isoas/us/overview.html**

[4] SAM: **http://www.prestospace-sam.ssl.co.uk**
PrestoSpace: **http://www.prestospace.org**

- **Persistent** identifiers [5] – the various efforts within web technology to counter the general tendency for resources on the web to “go missing” (broken link; dead link; link rot; 404 error). Estimates vary, but figures around 30% (almost regardless of the context) are common [6].

- **Persistent** resources [7] – the links or identifiers are a means to an end; persistence of the resources is the real issue.

Archives have exactly the same two concerns: not losing their metadata and not losing the archive contents. The metadata is just a way to get to the content – identifiers, descriptors and finding aids.

Digital archives will have files to hold the content, and a storage system using a sort of addressing scheme to locate the content. Thus, digital archives share the problems of the Worldwide Web and the IT industry, specifically the concern for persistence.

The storage industry does not use the term *persistence*. Typically, storage industry information relevant to losing stored data is expressed in terms of error rates (of the data-reading process), failure rates (at the device level) and media life expectancy.

There is a real gap here, because archivists have NO interest in read error rates and MTBF (Mean Time Between Failures), and it is a conjecture of this article that “digital archivists” will also have no interest in media life expectancy. Meanwhile the storage industry provides data about storage media systems, and NOT directly about the persistence of the content.

The author’s view is that the storage industry provides this sort information because that is the easy thing to do. They have information on media and systems, on its performance and failures. When the storage industry talks to archives, they should consider providing the information that archives really want: will the content still be there in 20 years? Will it persist?

Specifically, archivists want to know:

- **How much content will be lost**, every year for N years?
  This is the one figure an archive can use to decide whether or not a storage strategy is acceptable.

- **What is the statistical distribution of the probability of loss**?
  This information allows an archive to assess the degree to which performance (of the storage strategy) can be trusted. It’s no good investing in a strategy with a 1% projected loss, if there is a 50% chance that the loss can be 10 times higher. This may look a bit complex and exotic – statistics about statistics – but it’s exactly the same complexity of information an insurance company uses to compute life assurance premiums. Only when an archive knows the *confidence interval* around the probability of loss, can it make informed decisions about control of risk.

- **How do the probabilities vary with N (how do they vary over time)?**
  This is again basic information, because storage strategies need to be re-assessed regularly. It may well make sense to change strategy after a shorter rather than a longer time, because probability of loss may well increase over time (or the costs – of keeping losses from rising – may themselves rise). A good horse to bet on can, in time, turn into a tired horse or an expensive horse. We are all familiar with this situation, especially with respect to being a car owner. Consumer guides to car ownership provide relevant information. Archives would like the same sort of information from storage providers.

[5] Persistent identifiers: **http://www.nla.gov.au/padi/topics/36.html**

[6] Frank McCown, Sheffan Chan, Michael L. Nelson and Johan Bollen: **The Availability and Persistence of Web References in D-Lib Magazine**
Procs of the 5th International Web Archiving Workshop and Digital Preservation (IWAW-05), 2005.

[7] Persistent resources: **http://www.w3.org/Consortium/Persistence**

- **How do the probabilities vary with cost?**
  We all expect to "get what we pay for". We fully expect that a storage strategy with 99% persistence over 20 years would cost more than one with 95% persistence. How much more? Archives simply cannot get this information – not because the vendors won't say, but because the storage industry simply does not compute the statistics that the archivists most want to see.

Archives want to buy an insurance policy for their digital contents – if only they could get a quotation for the cost!

Currency

Currency is also not a standard term, as the terminology in the digital preservation area is still being established. The problem is format obsolescence, and currency refers to whether a storage strategy can deliver data as *usable* content – usable by current technology.

There is much work in digital preservation on format obsolescence. It is a recognized problem and much has been done to develop and implement solutions. For digital files in general, major institutions such as The National Archive in the UK and the US Library of Congress are developing software repositories [8] for legacy software. Many institutions are developing strategies to keep content usable (e.g. UKOLN in the UK [9], PADI [10] in Australia).

Persistence is a dimension of digital archive storage where an archive can expect the storage industry to come up with relevant statistics. The issue of *currency* is more difficult, but the whole digital library and digital preservation community has identified this problem and is working on solutions.

A digital broadcast archive has two main choices:

- Keep the original content as is, and ensure that there will always be players available to render the content into usable audio and video signals;
- Migrate the content as formats become obsolete.

The first option is fraught with problems, as it is an immense ambition to make players not only available, but to have those players where they are needed, namely alive and working on the desks of the archive users.

The second option is a chore, but one that keeps content viable. It the migration route is chosen, persistence is also affected – because updating files for currency requires that the files be read and re-written, which is a basic "refresh" operation that could well be a cornerstone of the strategy for persistence.

Ideally, the storage industry would supply information covering costs of such "refresh" operations, so that an archive could balance the benefits (for both persistence and currency) against costs of such a major operation as re-formatting an entire audiovisual collection. In practice, the storage industry does not supply this information – because it is about the use of devices rather than about the devices themselves, and so it is "the customer's business".

The *currency* issue takes precedence over the simple statistics that the storage industry does provide. What is the advantage (for anyone) of a medium that will hold a file for 100 years, or even 40 years, if the file format itself becomes unplayable within 10 years? As a reminder of the problem, what proportion of document or PowerPoint files from 1996 can be opened today? If it is less than 99.5%, then it is below the minimum persistence level likely to be required by archives.

[8] PRONOM: **http://www.nationalarchives.gov.uk/pronom/**
LOC Digital Formats: **http://www.digitalpreservation.gov/formats/**

[9] UKOLN: **http://www.ukoln.ac.uk/interop-focus/gpg/Preservation/**

[10] PADI: **http://www.nla.gov.au/padi/**

Storage service requirements: what storage does for archives

Storage may seem to be the central issue to a digital archive, but persistence and currency are the essentials. The remaining technical requirements are only two:

- Size of the storage;
- Bandwidth of the access to the storage.

Size is the easier of the two. The complications are decisions about file formats and the degree of compression to be used on master-quality files (if any – because archivists hate compression!). A "ready-reckoner" for calculating storage requirements is available on the PrestoSpace SAM website [11].

Storage size and file *persistence* are related. The more copies, the more redundancy within copies – the greater the chance a file will not disappear. But archives should not get involved in these complex interrelations.

Bandwidth has to do with the service requirements of the archive: how many users, how many concurrent users, where they are and how they are connected. It should be noted that for online archives with a large number of users (e.g. the general public), the bandwidth costs may far outweigh the storage costs – to the extent that the estimated storage cost could be less than the expected error in the estimated bandwidth costs, in which case the storage is effectively (or comparatively) free.

Storage media requirements

This article will say nothing about storage media requirements. The point of view so far has been that what matters is the service. If a storage service provider can pull together a service based on magnetic tape, or minidiscs, or surplus 8″ floppies from the 1970s, or holographic or molecular or optical tape storage – or even by bouncing data to the moon and storing it in the delay time – it simply doesn't matter to the archive. All that matters is that the storage is persistent, big enough, fast enough – and that when the files come back they can be used (currency). The fascination with storage media can be left to the storage industry. It's not the business of archivists.

The Future of digital archives

Digital archives are inevitable, and will solve certain problems:

- contention for the one tape that everybody wants at the same time;
- circulation control;
- chasing returns;
- making extra copies;
- getting material to distant places, quickly.

However, new issues will arise. The author foresees the following as new problems or at least new perspectives arising from digital archiving:

- no more storage in the archive – archives and storage parting company;
- different kinds of failure – analogue media failed "locally" and therefore partially whereas digital media tends to fail totally (when it fails);

[11] PrestoSpace: **http://www.prestospace.org**
Sam: **http://prestospace-sam.ssl.co.uk/**
Ready-Reckoner: **http://prestospace-sam.ssl.co.uk/tutorials/56/61.html**

- re-emergence of the bottom drawer – new ways for people to make private hoards rather than sending material for proper archiving and general use.

Archives and storage parting company

The claim has been made that as archives go digital, they begin to part company with storage. Three examples are: NRK in Norway, the broadcast and heritage archive B&G (National Institute of Sound and Vision) in The Netherlands and finally here in the BBC.

NRK [12]

Starting four years ago, the NRK sound archive entered into a joint project with the Norwegian national audio collection, to share the technology and costs on "going digital". Contents of the radio archives in Oslo were sent north of the Arctic circle 1100 km away for digitization, and for storage on a mass-storage system used jointly by NRK and the national archive. So, progressively, the physical contents of the radio archive have disappeared and reappeared online from a server 1100 km away. The delivery time into edit suites has been reduced, and there are added service advantages of online audition (rather than booking tapes from the archive).

B&G [13]

The Netherlands had its "big bang" earlier this year, with all TV material distributed from a digital playout server. The company providing this service, NOB, also provides an archive storage service to B&G. So no new material is coming to the B&G – it's being held for them by NOB. B&G continues to manage the cataloguing and access, and the fact that they "don't hold anything anymore" has not caused trauma and, in fact, has largely passed unnoticed. Who knows where electronic content comes from anyway? And who cares, so long as it comes?

BBC

The BBC outsourced its entire IT services to Siemens about two years ago. Storage will be provided by Siemens to the whole BBC, including archive storage. This arrangement is not seen by the archive as causing a problem – in fact it will be a relief – so long as we can have the *persistence* and *currency* we need, along with the delivery times and bandwidth our customers need. We do have issues about getting the statistics we need in order to satisfy ourselves about *persistence*. We want percentage-of-loss figures over 20 years – as a function of cost. As already stated, these are not off-the-peg figures in the IT world.

Missing technology: graceful failure

For decades, an error in reading a videotape has been subject to "concealment" – a line could be replaced by the contents of a previous line, allowing playback to continue. Clearly there is a limit to how much concealment can be tolerated, before it becomes visible and can hardly be called concealment. But a level of a few errors PER FIELD would be perfectly acceptable.

In the digital world, IT systems are designed to read out a file perfectly. There are levels of organization that underlie that capability, and it is all very impressive when it works. But when it fails, in the general case the whole file is rejected. Precisely what happens depends on the file-management system, operating system and individual application details – but it is common for a "cannot read file" or "cannot open file" message to appear, and nothing further can be done.

There is "file rescue" technology, but that is esoteric and in the hands of system managers, far away from the user of a digital archive who is having a problem. VTR concealment was right there, acting in real time exactly when and where needed, to keep things going.

[12] IASA: **htpp://www.iasa-web.org/Grimstad_digitalradioarchive.pdf**

[13] PrestoSpace: **htpp://www.prestospace.org/user_group/20060518/Memo-PS_Preservation_Factory_Workshop_18-05-06.pdf**

It may be a pipe dream, but it would certainly be advantageous to the broadcasting industry to have media files that were error-tolerant. At the basic level, something as simple as making a block of data storage (on a disc or tape) equivalent to an exact number of video lines ("1" would be a very good number) – would allow a "bad block" to be ignored, the previous block to be used in its place (concealment reborn) so that the user can carry on.

The BBC will pursue some of these ideas under DTI-supported research starting this autumn [14].

The "bottom drawer"

One of the problems with an archive inside a larger business, as is common in broadcasting, is that the IT definition of "archive" is spreading. Edit and other post-production systems, and playout systems, are now commonly server-based, as IT applications. Data that does not fit on the server can then be "archived", following the IT model.

The problem with these application-specific archives is just that: they are "application-specific", typically using proprietary methods. Even if the audiovisual content with these "IT archives" is on non-proprietary formats, the metadata is invariably in a proprietary database, or the link between the metadata and the content is in the proprietary database. **There is no agreed and widely-implemented standard for the storage component of asset management, edit and playout systems**. It is the broadcast archive that provides a common format for metadata and media across the business, and application-specific IT archives are resurrecting the "bottom drawer problem".

A broadcaster benefits from the re-use of assets, and archives were set up to enable that re-use. The enemy of the archive has always been the local "bottom drawer", where material was kept for re-use by one person or unit within the business. It was inaccessible to all others – who couldn't even find out about the contents of the bottom drawer, and so could not even try to use those contents.

Now IT archives are creating application-specific bottom drawers, and the advances in storage mean that a great deal of material can be squirreled away in such archives at relatively low cost. The cost matters, because if the costs were high, these private archives would not be so likely to occur.

A primary effort of broadcast archives is to create a single digital repository, serving the whole business, so that digital assets can be used across the business. This laudable task is frustrated at every turn by the digital bottom-drawer phenomenon. The archive – the true, business-wide archive – has to establish a way to get material out of every "nook and cranny" that is capable of having its own IT archive. Here again, cost is an issue: the application may come with IT-archive functionality built-in, or available at relatively low cost. The true archive needs to bid for funding to do something harder and more expensive: get the content and the metadata out of the digital bottom drawer, and get it into generic media and metadata formats that are common across the business. Worse still, the true archive has to try to do this for all the relevant applications. Adoption of web-service middleware and other general standards across the business may ease the cost, but it is still generally more expensive to build a true archive than to have a set of application-specific IT archives.

If the additional funding isn't forthcoming, the whole broadcast archive business will sink back into the bottom-drawer culture that was present when archives were first starting. Decades of effort and progress will be in danger of being undermined or lost, and the business will lose its ability to share and re-use assets (except within local units sharing a bottom drawer).

[14] Department of Trade and Industry

Richard Wright was educated at the University of Michigan (USA) and Southampton University (UK), and holds a B.Sc. in Engineering Science (1967), an M.A. in Computer Science (1972) and a Ph.D. in Digital Signal Processing – Speech Synthesis (1988).

Dr Wright has worked in acoustics, speech and signal processing for US and UK Government research laboratories (1968-76), at the University College of London (1976-80; Research Fellow) and at the Royal National Institute for the Deaf (1980-90; Senior Scientist). He was the Chief Designer at Cirrus Research from 1990 to 1994 (acoustical and audiometric instrumentation).

Richard Wright has been the Technology Manager of BBC Archives since 1994. He was also the Head of EBU working group, P/FRA – Future Radio Archives, and of the EC-sponsored project PRESTO (Preservation Technology).

Conclusions

Broadcast archives will digitize and move to mass storage – it's already happening across broadcasting. But the more we rely on IT technology for archive functionality, the more we will need a common language. We have an urgent need for ***persistence*** information from the storage industry.

Storage providers who understand the needs of broadcast archives and who can provide convincing statistics will be best placed to capture the archive storage market. Broadcasters look forward to leaving the storage-technology issues to the experts – as soon as the storage industry has developed true archive storage expertise, with convincing and comprehensive services.